Crime Data

17 datasets: (UCR FBI yearly crime data; 2003-2018, for 2004 and prior, offenses are broken down by cities and towns above and below 10,000; however, for 2004 specifically, only crime values for cities and towns greater than 10,000 are provided by the FBI.)

These datasets included values for which the FBI had caveats. These caveats were in the form of footnotes, and unfortunately footnotes didn't maintain the same pattern from year-year. Thus, one year "3" might correspond to the footnote for the FBI finding an over-reporting of values from a police agency, and another year "5" might coincide with that caveat. These caveats range from police departments under- or over-reporting crime statistics to data collection methodology for reporting of forcible rape not complying with UCR guidelines to a police agency's changing in reporting practices making them incomparable to previous years. Much of these footnotes coincided with NAs; for instance, the FBI did not report violent crime values for those police agencies that did not comply with UCR guidelines for methodology of reporting forcible rape. (As an example, in 2012, Chicago, Illinois and Minnesota did not use the same collection methodology complying with the "National UCR guidelines." In such instances, the FBI chose not to report violent crime values.

We scrutinized values with caveats at the state level which were mostly where FBI reports that changes in reporting practices by agencies made statistics incomparable to previous years. For instance, in 2012, when we looked at several values for New York State including Troy, New York City, and Schenectady, we found them to be comparable to previous years' values. For instance, in 2011 Schenectady's violent, property, and total crime were respectively 633; 3,098; and 3,731; and in 2012, they were 626; 2,828; 3,454. We investigated this on a broader scale by filtering values for each specific year's footnote corresponding to incomparable values using regular expressions. We then compiled all 17 datasets with values which were deemed by the FBI to be incomparable to previous years' data. We z-scored all crime values (violent, property, and total) such that when we merged this dataset with the main dataset we compiled for all crime values over 2003-2018 period, we could then lag all crime values by one year to look at each place's standardized yearly change in crime values. We found that marked values by the FBI and previous years did not differ considerably and that, in many instances, larger differences occurred between other years for these places. Additionally, we found that the distributions in differences among the marked FBI year and previous years was similar to the overall distribution of differences among previous years. The distribution for FBI marked values that were incomparable to previous years had a mean of 0 and a SD of 1.05, while the distribution for non-marked values had a mean of zero and a SD of .95

Thus, we had a total of 136,118 values. We did a check to ensure that there were no repeated towns/cities (that all of these were distinct values for each year). There was only one non-distinct value of the 136,118 observations. By creating a dataset of distinct values (136,117) and anti-joining that dataset with the full dataset (116,118 values), we located this value to be Prescott, Arizona in 2003. Looking at the 2003 datasets (for both greater than and less than 10,000 people, we found that the error was in the less than 10,000 people dataset, for which Prescott, AZ should not have been included by the FBI). We simply removed this value.

District finance data

16 datasets (Census, Annual Survey of School System Finances: 2003 through 2018)

We simply compiled these datasets into one larger dataset. We later imposed checks on the validity of these values which we discuss below.

Merging Crime Data with District Finance Data

13 datasets (NCES Geographic Relationship Files: places, 2013-2018; county subdivisions, 2015-2018; Missouri Mable: places 2000, 2010; county subdivisions 2000).

NCES created six datasets that merge places to school districts 2013-2018, and 4 datasets for county subdivisions (2015-2018). Missouri Mable, which is now maintained by University of Michigan, ICPSR, provides the possibility to create 2000 and 2010 datasets merging places with school districts; only their 2000 dataset works for county subdivision to school districts. Here, we used the 2000 Mable dataset for years up through 2009, and we used the 2013 NCES places dataset for years 2010-2013. For county subdivision values, we used the 2000 dataset for years 2000-2010 and the 2015 NCES county subdivisions dataset for 2011-2015. Because of the consolidation of districts over the years, the 2000 Mable datasets would have more school districts than in reality; however, since a district that has become consolidated districts would no longer exist, these values would have no counterpart in school finance datasets. We felt that this was the most conservative approach.

First, we merged district finance data with the corresponding places and county subdivision datasets above such that we had one places to districts dataset and one county subdivisions to districts dataset.

When two cities of the same name exist in one state, UCR data provides county values. An instance of this is "Adams" which exists in both Butler and Cambria counties in Pennsylvania. If county was provided we placed it into a separate dataset. We then merged these datasets separately with the above "places to districts" dataset, such that we had the larger dataset (without county names) and the smaller dataset (with county names, where multiple cities/towns of the same name exist in the same state) to look at which values failed to join.

In the larger dataset (without county names), we found that 38,387 observations (at this point, these are not distinct city/town values) failed to join. We then used the "anti-join" function in dplyr (Wickham), such that we could identify which values failed to join and discern patterns for why these values failed to join. Much of this was caused by epithets like "township", "borough," etc. in district data not coinciding with UCR FBI data. Thus, we stripped these epithets. After cleaning, 1,003 values joined, while 37,772 did not. We gained 25,677 values by trimming white space around the "name" variable (for cities/towns) In our smaller dataset, in which county names were provided, 1500 observations did not join. Our hypothesis was that these values were actually county subdivisions, which is why they failed to merge with our places dataset (It's worth noting here that this was an iterative process and that we only created a "county subdivisions to districts" dataset after noticing a large number of observations that failed to join). Importantly, here, there would be no doubling up for places which are also county subdivisions because the only values which were merged with the "county subdivisions to districts" dataset were specifically those values which did not join with the "places to districts" dataset.

We used the Missouri Mable 2000 dataset with county subdivisions to create a crosswalk between counties and county subdivision identifiers such that we could merge the UCR crime values for which counties were provided (we reclaimed 1,223 of the 1500 values that did not join). Additionally, of the 37,772 values that did not join, we reclaimed 15,485 of these values from county subdivisions).

The second issue we had to contend with was the fact that while cities/towns of the same name could exist within one state in district datasets, if all of these cities/towns didn't report

crime data (and were accounted for with county names), then these crime values would join with all cities/towns of the same name in one state (since crime data were only really qualified with city and state information in UCR datasets). This would be a serious issue; given that ultimately, if, for instance, Springdale Police Department in Lexington County of South Carolina reported crime but Springdale Police department in Lancaster county of South Carolina did not, then these values would be conflated. We dealt with this issue by finding the number of cities/towns in the crosswalk dataset (places to districts) with distinct "place_id" observations (25,088); and the number of distinct "city/town" and "state" observations (25,060). The difference between these values was the number of observations that would be problematic and the city/town values identified in this process were the names of towns/cities that were doubled up within a state. We created a dataset with these specific city/town and state values and filtered them from the places-districts dataset.

We repeated the same process for county subdivisions and found that repeated names were much more prevalent among county subdivisions and had to remove a total of 6,471 distinct cousub_ids.

In downloading the 2000 Mable crosswalk dataset for county subdivisions, we found that we needed to create these files in rough collections of 10 states per request, otherwise the system was unresponsive. One must create Place (7 digits) or county subdivision ids (10 digits) by using a combination of state and county identifiers.

Population data

4 datasets (Decennial census: 2000, 2010; Population Estimates Program)

We then added population data to our main dataset. The first two datasets we used were decennial Census data. The second two datasets were from the Census Population Estimates Program for years 2001-2009 and 2011-2018. We extracted both place (filtering by summary level 162) and county subdivision information from these datasets (filtering by summary level 061) information from all four datasets. Thus we had one dataset for merging "places" (using place_id, name, and state) and another dataset for merging county subdivisions (by cousub_id, name, state). We then combined these datasets to have one main dataset.

Proportion Children in Poverty

16 datasets (Small Area Income and Poverty Estimates, SAIPE, Program: 2003-2018)

SAIPE offers yearly datasets for proportion of children (k-12) living in poverty for both counties and school districts. We used school district data, here, given that these estimates occur on a more granular level than counties and hence, yield greater fidelity (4x the number of school districts as counties). These datasets were joined to our main dataset by year and school district id. An explanation of the uncertainty of these estimates can be found here: https://www.census.gov/programs-surveys/saipe/guidance/district-estimates.html

Presidential Elections

1 dataset. (MIT Election Data + Science Lab: 2000-2016)

Implementing presidential elections, given that they only occur every four years, involved making a choice regarding the preservation of values. We split the county data for each election over four years (the one year before the election and the three years following the election: such that the 2004 election would span the years 2003:2006, the 2008 election would span the years 2007:2010, etc.). Our logic was that this would best capture any changing of

views over time (whereas considering the results over the four years following the election would suggest that political views changed only after the election).

Law Enforcement Datasets

30 datasets (US Department of Justice and FBI, LEOKA police employee data: 2003:2016, not yet published for 2017 and after; and UCR FBI, police employee data: 2003-2018).

We joined the LEOKA (from ICPSR) and FBI datasets by town/city by state by year. We then joined these datasets to our main dataset using the same methods as those used to join crime data (filtering out epithets the following epithets: "village| town| borough| boro| and| regional| resort| county| metropolitan." The Bureau of Justice does provide a crosswalk between ORI (originating agency identifiers) codes and places; however, we found this dataset to be problematic and found that we preserved more values with our method (furthermore, the crosswalk only pertains to monthly crime data, which is mainly published as raw, where as the final yearly values have been processed).

From here, we then filtered our main dataset by those values with SCHLEV = "03" (school district values) such that we were left only with unified school districts. We also filtered out any school districts with enrollment below 100. At this point, we checked those district spending values still included in the dataset by lagging district values by one year such that we could see whether year-year values had increased by over $5000, such that we could flag these values, investigate and possibly replace them with more appropriate values from state Department of Education or district websites. (We replaced 59 values.)

Joining Other Covariates

The Census stopped asking all individuals for income and educational attainment after 2000. From that point forward, the American Community Survey (ACS) began surveying random individuals. For places with fewer than 65,000 individuals, these estimates are produced yearly but are averaged over a five-year period. We used the averaged estimates (in most cases, these existed starting in 2009-2010) for every year.

We obtained both Place_id and Cousub_id datasets for all of the following variables. Given that the following datasets were generally not available on a yearly basis until 2009 or 2010, from American Community Survey (ACS), we spread these years out as well, similar to how we dealt with presidential elections. We joined datasets below sharing exact years, such that education attainment, median income, and housing costs were joined; per capita income and unemployment were joined, and race remained its own set of datasets. For education attainment, median income, and housing costs we spread the 2000 dataset from 2003:2005; the 2010 dataset from 2006:2010; and all other datasets assumed their appropriate year. Then for per capita income and unemployment datasets we spread the 2000 dataset over the years 2003:2005; and the 2009 dataset from 2006:2009; and then all other datasets assumed their appropriate year. Lastly, we had race variables for the years 2000, 2010, and 2015: we distributed the 2000 dataset over the years 2003:2007; the 2010 dataset over the years 2008:2013; and the 2015 dataset over the years 2014:2018

Education attainment

20 datasets (Decennial: 2000; ACS 5-year estimates: 2010-2018)

Median Income
20 datasets (Decennial: 2000; ACS 5-year estimates: 2010-2018)

Per capita income
22 datasets (Decennial: 2000; ACS 5-year estimates: 2009-2018)

Unemployment
22 datasets (Decennial: 2000; ACS 5-year estimates: 2009-2018)

Housing costs
20 datasets (Decennial: 2000; ACS 5-year estimates: 2010-2018)

Race
6 datasets (Decennial: 2000; ACS 5-year estimates: 2010, 2015)

We then reduced our dataset such that for any cities/towns overlapping two counties, the dataset would only retain the observation for which the population was the largest.

For all independent variables that we logged, we smoothed zeros to be to .000001 and then logged and centered our independent variables. We then grouped our dataset by "place_id" (which includes county subdivision identifying values) and summarized by the mean such that the same place couldn't have observations in multiple districts (we did place_id as opposed to "city" because place_ids (and cousub IDs) are unique while town and city names are not). Ideally, this variation could be handled in a GLMM model, but it can be problematic in cases where there is a lot of overlap (mixed effects) among places and school districts. Thus, our final dataset only had one identifier (place/cousub_id per year (we also looked at range by city/town and county and this is where we would've summarized by range for those datasets, vs the mean).

From here we imputed via multivariate imputations by chained equations. We created prediction matrices to maintain the multilevel structure of the data in our imputations. We used the MICE package (van Buuren); we imputed 5 values for each missing value, employing 15 iterations; we then averaged these values using the merge_imputations() feature of sjMisc (Ludecke).